# A New Approach to Inference in Multi-Survey Studies with Unknown Population Size

Kyle Vincent*and Saman Muthukumarana†

December 1, 2015

**Abstract**

We investigate a Poisson sampling design in the presence of unknown selection probabilities when applied to a population of unknown size for multiple sampling occasions. The fixed-population model is adopted and extended upon for inference. The complete minimal sufficient statistic is derived for the sampling model parameters and fixed-population parameter vector. The Rao-Blackwell version of population quantity estimators is detailed. An application is applied to an emprical population. The extended inferential framework is found to have much potential and utility for empirical studies.

Keywords: Complete minimal sufficient statistic; Design-based inference; Mark-recapture; Population total; Poisson sampling; Rao-Blackwell theorem.

The authors wish to thank Chris Henry, Kim Huynh, Richard Lockhart, Louis-Paul Rivest, Michael Stephens, and Steve Thompson for helpful suggestions on the preparation of this manuscript. The authors also wish to thank John Potterat and Steve Muth for making the

*Currency Department, Bank of Canada, 234 Laurier Avenue West, Ottawa, Ontario, CANADA, K1A 0G9, *email*: kvincent@bankofcanada.ca

†Department of Statistics, University of Manitoba, 338 Machray Hall, Winnipeg, Manitoba, CANADA, R3T 2N2, *email*: Saman.Muthukumarana@UManitoba.CA

Colorado Springs data available. All views expressed in this manuscript are solely those of the authors and should not be attributed to the Bank of Canada.

# 1 Introduction

There is an abundance of mark-recapture literature on the estimation of population size under many different settings and sampling models. In such settings there is usually an absence of a sampling frame, like in the study of a large, wildlife, or hidden population, thereby not permitting the investigator full control over the sample selection procedure. Hence, selection probabilities are likely to be unknown and complications for inference of quantities like the population total will arise. Estimation procedures for these quantities has received little attention in the literature. In this paper we address the need for such efficient estimation procedures. We consider the Poisson sampling design as such studies can naturally fall under a framework that bases inference on this design.

The focus of this manuscript is twofold. First, we establish a foundation for inference of population quantities when Poisson sampling is used and multiple samples are selected. We extend on the classic case in the survey sampling setup by adopting the usual framework based on a design-based approach to inference; we base inference on the fixed-population model and consider how it changes when the population size is unknown and hence there is an absence of a sampling frame. Our strategy for inference therefore relies on a combination of classic survey sampling inference techniques and mark-recapture concepts. Second, we demonstrate the utility of the Rao-Blackwell theorem in survey sampling studies. In particular, we show that if it can be assumed that groups of individuals share the same selection probability when a Poisson sampling design is used, so much so that even if it effects a large number of strata over the population, then there is much potential for improving on preliminary estimates with the Rao-Blackwell theorem under the design-based approach to

inference.

Traditional design-based inference in survey sampling rests on the use of a fixed-population model where the population size is known; see Cassel et al. (1977) for details. Most research based on the fixed-population model considers the case where a sample is selected with an ignorable design; such a design is defined as one in which selection probabilities are known and the design does not depend on response values outside the sample. Basu (1969) derived the minimal sufficient statistic for the population parameter vector when sampling is based on such a design, and Cassel et al. (1977) later showed that it is not complete. In contrast to the traditional approach, the Poisson sampling design in our study is based on selection probabilities that are unknown and is therefore non-ignorable. We show that this results in a different minimal sufficient statistic than that seen in the classic setting, and that it enjoys the property of being complete.

Two sources of research that are relevant to the work presented in this manuscript are detailed as follows. First, Kindahl (1962) investigated the use of simple and stratified random sampling designs when the population size is unknown. In his article, he bases estimates for the population total on a combination of expansion estimators for population stratum totals; each expansion estimator is based on the product of estimates of a corresponding stratum size and mean where estimates for the stratum sizes rely on classic mark-recapture concepts. Similarly, we exploit mark-recapture concepts at the inference stage, but in contrast they are used to assist in obtaining estimates of the selection probabilities as these are plugged into Hansen-Hurwitz and Horvitz-Thompson type estimators, much like in the approach used by Fattorini (2006).

Second, Ahmad et al. (2000) derive the complete and minimal sufficient statistic for the population size and total of responses of interest when sampling is based on a sequential sampling design that terminates when a predetermined number of repetitions occur. Similarly, we derive the complete and minimal sufficient statistic but now for the population

parameter vector when a Poisson sampling design is used. Hence, estimators with uniformly lowest mean square error can be obtained for any function of the population responses with the Rao-Blackwell theorem.

The paper is organized as follows. Section 2 introduces the sampling design and notation. Section 3 provides the derivation of the complete and minimal sufficient statistic. Section 4 outlines the Rao-Blackwellization procedure. Section 5 presents the population total estimators and variance estimators. Section 6 presents the results from a simulation study based on an empirical population. Section 7 concludes the manuscript with a discussion and direction for future work.

# 2 Sampling Design and Notation

Define $N$ to be the size of the population and $K$ to be the number of samples obtained for the study. Define the number of stratum in the population to be $G \leq N$ and $\underline{p} = (p_1, p_2, ..., p_G)$ to be the sampling model parameter vector of distinct selection probabilities where each entry corresponds with a stratum, as follows. Define $P_{ik}$ to be the probability that unit $i$ is selected for sample $k$ where $i = 1, 2, ..., N$, and $k = 1, 2, ..., K$. Now, suppose unit $i$ belongs to stratum $j$, where $j = 1, 2, ..., G$. Define $P_{ik} = p_j$ so that selection probabilities only depend on stratum membership, that is selection probabilities are homogenous both within strata and over all sampling occasions. To clarify, all units in stratum $j$ have an equal probability $p_j$ of being obtained on each sampling occasion.

Following the traditional fixed population approach to inference in sampling (see Cassel et al. (1977) and Thompson and Seber (1996) for further details) the typical objects of inference are $\underline{\theta} = (y_1, y_2, ..., y_N)$ where $y_i$ refers to the responses of interest (including stratum membership) of unit $i$, and now also the sampling model parameter vector $\underline{p} = (p_1, p_2, ..., p_G)$ since these are unknown. The *original data* observed for the study is $d_0 = ((s_k, \underline{y}_{s_k}) : k = 1, 2, ..., K)$

where $s_k$ refers to the units selected for sample $k$ and $\underline{y}_{s_k}$ is the corresponding vector of responses of interest of the units selected for sample $k$.

We shall clarify the notation we have introduced for our sampling design and notation setup with the following example. Suppose that $K = 3$ samples are obtained from a population with $G = 2$ stratum. Consider the outcome presented in Table 1 where letters refer to units and numerical subscripts denote the stratum the unit belongs to. In this example only stratum memberships are observed on sampled units and hence the original data can be expressed as $d_0 = (((A,1),(B,1)),\ (C,2),\ ((A,1),(C,2))$.

Table 1: An example of data observed from a multiple survey study.

| Original Data | | | |
|---|---|---|---|
| Sample 1 | $A_1$ | $B_1$ | |
| Sample 2 | | | $C_2$ |
| Sample 3 | $A_1$ | | $C_2$ |

# 3 Minimal Sufficiency and Completeness Results

Define $r_d(d_0) = d_R = ((s, \underline{y}_s), \underline{C})$ where $r_d$ is the *reduction function* that maps the original data to the *reduced data* $d_R$, $s = \cup_{k=1}^{K} s_k$, and $\underline{C} = (C_1, C_2, ..., C_G)$ where $C_j$ is the total number of selections made from stratum $j$ over all samples. For example the reduced data corresponding with the observed data presented in Table 1 is $d_R = (((A,1),(B,1),(C,2)),\ (3,2))$.

Following the traditional approach to inference in survey sampling (see Cassel et al. (1977) and Thompson and Seber (1996) for details) we will make the definition that a parameter vector $\underline{\theta}$ is *consistent* with a data value $d_R = ((s, \underline{y}_s), \underline{C})$ if $\underline{\theta}$ can be partitioned and sorted such that the first $n = |s|$ unit labels and corresponding observed $y$-values of $\underline{\theta}$ coincide with $s$ and $\underline{y}_s$, respectively. For all $d_R$ define $\Theta_{d_R}$ to be the subset of all $\underline{\theta}$ that are consistent

with $d_R$. For example, parameter vectors that are consistent with the reduced data from the previous example, $d_R = (((A,1),(B,1),(C,2)),\ (3,2))$, are:

$\underline{\theta} = (y_A = 1, y_B = 1, y_C = 2), (y_A = 1, y_B = 1, y_C = 2, y_{\text{“4”}} = 1), (y_A = 1, y_B = 1, y_C = 2, y_{\text{“4”}} = 2), (y_A = 1, y_B = 1, y_C = 2, y_{\text{“4”}} = 1, y_{\text{“5”}} = 1), (y_A = 1, y_B = 1, y_C = 2, y_{\text{“4”}} = 2, y_{\text{“5”}} = 1)$,..., etc.

It shall be understood that permutations of parameter vectors are equivalent. For example, $(y_A = 1, y_B = 1, y_C = 2) \equiv (y_B = 1, y_C = 2, y_A = 1)$.

Notice that, unlike the traditional fixed-population model setup, as $N$ is unknown the unit labels are not necessarily indexed as $1, 2, ..., N$. In fact, in our setup, and by the definition presented in Cassel et al. (1977), the units are *not identifiable* since the labels of units are not known. However, we shall make it clear that each (repeated) selection can be identified in the sense that they and their corresponding variables of interest are made known to the observer. Furthermore, for the purposes of inference, the units are arranged in a non-numerical/non-ordinal fashion, hence the use of the notation “4”, “5”, etc... for unobserved unit labels in the previous example.

Similar to how the likelihood for the unobserved data corresponding with the traditional survey sampling setup is expressed, the likelihood function for $\underline{\theta}$ and $\underline{p}$ given a specific realization of $D_0 = d_0$ in our setup is expressed as

$$L(\underline{\theta}, \underline{p} | D_0 = d_0) = P_{\underline{\theta},\underline{p}}(D_0 = d_0) = \prod_{j=1}^{G} \left[ \left( \frac{p_j}{1 - p_j} \right)^{C_j} (1 - p_j)^{KN_j} \right] I[\underline{\theta}\ \epsilon\ \Theta_{d_R}] \qquad (1)$$

where $N_j = \sum_i I_{\underline{\theta}}[y_i = j]$ is the size of stratum $j$. In the traditional survey sampling setup, that is where the population size is known and samples are selected via an ignorable sampling design, the likelihood of the unobserved values is flat. In contrast, the likelihood of the unobserved values and population parameters given in Expression (1) is not flat. This is a direct consequence of the non-ignorability feature of the Poisson sampling design when

selection probabilities and population size are unknown.

**Theorem**: The reduced data $D_R = r_d(D_0)$ is the minimal sufficient statistic for $(\underline{\theta}, \underline{p})$.

**Proof**: For any $d_0$, $\underline{\theta}$ and $\underline{p}$,

$$P_{\underline{\theta},\underline{p}}(D_0 = d_0) = \prod_{j=1}^{G} \left[ \left( \frac{p_j}{1-p_j} \right)^{C_j} (1-p_j)^{KN_j} \right] I[\underline{\theta} \, \epsilon \, \Theta_{d_R}] = T(\underline{\theta}, \underline{p}, d_R) t(d_0) \tag{2}$$

where $t(d_0) = 1$. By the Neyman-Factorization Theorem $d_R$ is sufficient for $(\underline{\theta}, \underline{p})$.

To show the minimality of the claim, we will make use of the theorem which gives the following. If $\underline{X}$ is a sample with probability mass/density function $f(\underline{X}|\underline{\phi})$ and $V(\underline{X})$ is a function of $\underline{X}$ such that, for any two sample points $\underline{x}$ and $\underline{x}^*$, for some function $h(f(\underline{x}), f(\underline{x}^*)) > 0$ $h$ is constant as a function of $\underline{\phi}$ if and only if $V(\underline{x}) = V(\underline{x}^*)$, then $V(\underline{X})$ is the minimal sufficient statistic for $\underline{\phi}$ (Casella and Berger, 2002).

Now, take any $d_0 = ((s_k, \underline{y}_{s_k}) : k = 1, 2, ..., K)$ and $d_0^* = ((s_k^*, \underline{y}_{s_k^*}) : k = 1, 2, ..., K)$ where $P(d_0) > 0$ and $P(d_0^*) > 0$. Let $r_d$ be the usual reduction function where $r_d(d_0) = d_R = ((s, \underline{y}_s), \underline{C})$ and $r_d(d_0^*) = d_R^* = ((s^*, \underline{y}_{s^*}), \underline{C}^*)$. Suppose that

$$P_{\underline{\theta},\underline{p}}(D_0 = d_0) = h(d_0, d_0^*) P_{\underline{\theta},\underline{p}}(D_0 = d_0^*) \tag{3}$$

where $h(d_0, d_0^*)$ is independent of $(\underline{\theta}, \underline{p})$. Then we can re-express (3) as

$$\prod_{j=1}^{G} \left[ \left( \frac{p_j}{1-p_j} \right)^{C_j} (1-p_j)^{KN_j} \right] I[\underline{\theta} \, \epsilon \, \Theta_{d_R}] \tag{4}$$

$$= h(d_0, d_0^*) \prod_{j=1}^{G} \left[ \left( \frac{p_j}{1-p_j} \right)^{C_j^*} (1-p_j)^{KN_j^*} \right] I[\underline{\theta} \, \epsilon \, \Theta_{d_R^*}]. \tag{5}$$

As the probability of obtaining the original data is greater than zero and $h(d_0, d_0^*)$ does not depend on $\underline{\theta}$ or $\underline{p}$, the indicators must be zero or one at the same time. Therefore, $\Theta_{d_R} = \Theta_{d_R^*}$

so it must be that $s = s^*$ and $\underline{y}_s = \underline{y}_{s^*}$. Furthermore, the equality

$$\prod_{j=1}^{G} \left( \frac{p_j}{1-p_j} \right)^{C_j} = \prod_{j=1}^{G} \left( \frac{p_j}{1-p_j} \right)^{C_j^*} \tag{6}$$

must hold for all values of $\underline{p}$. This only happens if $C_j = C_j^*$ for all $j = 1, 2, ..., G$; the rationale for this can be provided by taking the logarithm of both sides of (6) and then setting the difference of the two sides equal to zero. Therefore, it must be that $d_R = d_R^*$ and hence $D_R$ is the minimal sufficient statistic for $(\underline{\theta}, \underline{p})$. □

**Theorem**: The statistic $D_R = r_d(D_0)$ is complete.

**Proof**: Choose any measurable function $g$ that is independent of $(\underline{\theta}, \underline{p})$. Suppose that

$$E_{\underline{\theta},\underline{p}}[g(D_r)] = \sum_{D_r = d_r} \left( g(d_r) P_{\underline{\theta},\underline{p}}(D_r = d_r) \right) = 0. \tag{7}$$

Index all possible $d_r$ as $d_r^{(a)}$ where $a = 1, 2, ..., A$. Now, suppose that

$$\begin{aligned} E_{\underline{\theta},\underline{p}}[g(D_r)] &= \sum_{a=1}^{A} \left( g(d_r^{(a)}) P_{\underline{\theta},\underline{p}}(D_r = d_r^{(a)}) \right) \\ &= \sum_{a=1}^{A} \left( g(d_r^{(a)}) \prod_{j=1}^{G} \left[ \left( \frac{p_j}{1-p_j} \right)^{C_j^{(a)}} (1-p_j)^{KN_j} \right] \right) \\ &= \prod_{j=1}^{G} \left[ (1-p_j)^{KN_j} \sum_{a=1}^{A} \left( g(d_r^{(a)}) \prod_{j=1}^{G} \left( \frac{p_j}{1-p_j} \right)^{C_j^{(a)}} \right) \right] \\ &= 0. \end{aligned} \tag{8}$$

As $N_j > 0$ and $0 < p_j < 1$ for all $j = 1, 2, ..., G$, and hence the product terms are greater than zero, it must be that $g(d_r^{(a)}) = 0$ for all $a = 1, 2, ..., A$. Hence, $P(g(D_r) = 0) = 1$ for all $(\underline{\theta}, \underline{p})$. Therefore, $D_r$ is a complete statistic. Furthermore, this reinforces the claim that $D_R$ is the minimal sufficient statistic. □

# 4 Rao-Blackwellization

Rao-Blackwellization is a mathematically powerful technique that can be used to improve the precision of an estimator. For cases involving the usual survey sampling setting this theorem has received much attention; see Thompson (2006) and Chao et al. (2011) for examples. For a brief discussion on and other examples of how to make use of the minimal sufficient statistic through the Rao-Blackwell procedure in the usual survey sampling setting, see Thompson (2012).

## 4.1 Point estimation

Recall that the reduced data in our setup to inference is $d_r = ((s, \underline{y}_s), \underline{C})$. A reordering of the original data is consistent with the reduced data if the following two observations hold. First, all $n = |s|$ members of $s$ are hypothetically selected for at least one of the $K$ samples. Second, a total of $C_j$ selections are made from $s$ and within stratum $j$ over all hypothetical samples. For example, recall that the reduced data of the original data illustrated in Table 1 is $d_R = (((A, 1), (B, 1), (C, 2)),\ (3, 2))$. Table 2 presents two reorderings of the original data that are consistent with the reduced data. Notice that sample reorderings consistent with the reduced data do not require the sampled units to be selected equally often as they were in the original ordering. Furthermore, also notice that members from different strata can move between samples simultaneously to give rise to consistent sample reorderings.

Table 2: Two data reorderings that are consistent with the reduced data of the observed data presented in Table 1. Letters refer to units and numbers refer to stratum memberships.

| Reordered data example 1 | | | |
|---|---|---|---|
| Sample 1 | $A_1$ | | $C_2$ |
| Sample 2 | | $B_1$ | $C_2$ |
| Sample 3 | $A_1$ | | |
| **Reordered data example 2** | | | |
| Sample 1 | $A_1$ | | |
| Sample 2 | | $B_1$ | $C_2$ |
| Sample 3 | | $B_1$ | $C_2$ |

In order to fully exploit the minimal sufficient statistic, Rao-Blackwellized/improved estimators can be obtained as follows. Upon obtaining $d_0$, define $\mathcal{R}$ to be the set of all reorderings of the original data that are consistent with the reduced data. Let $\hat{\gamma}_0$ denote the preliminary estimate depending on $d_0$ of a population quantity $\gamma$. For each hypothetical reordering $i \,\epsilon\, \mathcal{R}$ define $d_0^{(i)}$ to be the corresponding reordered sample data, $\hat{\gamma}_0^{(i)}$ to be the preliminary estimate obtained with reordering $i$, and $C_{j,k}^{(i)}$ to be the number of individuals from stratum $j$ that are selected on sampling occasion $k$ under reordering $i$. The Rao-Blackwellized version of the preliminary estimator $\hat{\gamma}_0$ is

$$\begin{aligned}
\hat{\gamma}_{RB} = E[\hat{\gamma}_0|d_r] &= \sum_{i\epsilon\mathcal{R}} \left( \hat{\gamma}_0^{(i)} p(d_0^{(i)}|d_r) \right) = \frac{\sum\limits_{i\epsilon\mathcal{R}} \left( \hat{\gamma}_0^{(i)} p(d_0^{(i)}) \right)}{\sum\limits_{i\epsilon\mathcal{R}} p(d_0^{(i)})} \\
&= \frac{\sum\limits_{i\epsilon\mathcal{R}} \left( \hat{\gamma}_0^{(i)} \prod\limits_{k=1}^{K} \left\{ \prod\limits_{j=1}^{G} \left[ p_j^{C_{j,k}^{(i)}} (1-p_j)^{N_j - C_{j,k}^{(i)}} \right] \right\} \right)}{\sum\limits_{i\epsilon\mathcal{R}} \left( \prod\limits_{k=1}^{K} \left\{ \prod\limits_{j=1}^{G} \left[ p_j^{C_{j,k}^{(i)}} (1-p_j)^{N_j - C_{j,k}^{(i)}} \right] \right\} \right)} = \frac{\sum\limits_{i\epsilon\mathcal{R}} \left( \hat{\gamma}_0^{(i)} \prod\limits_{j=1}^{G} \left[ p_j^{C_j} (1-p_j)^{N_j - C_j} \right] \right)}{\sum\limits_{i\epsilon\mathcal{R}} \left( \prod\limits_{j=1}^{G} \left[ p_j^{C_j} (1-p_j)^{N_j - C_j} \right] \right)} \\
&= \sum_{i\epsilon\mathcal{R}} \hat{\gamma}_0^{(i)} / |\mathcal{R}|. \qquad (9)
\end{aligned}$$

Notice that the estimator presented in (9) does not depend on $(\underline{\theta}, \underline{p})$, reinforcing the claim that $d_r$ is a sufficient statistic for $(\underline{\theta}, \underline{p})$. Further, all data reorderings have an equal probability of being obtained.

## 4.2 Variance estimation

To estimate the variance of the improved estimator, the decomposition of variances gives

$$\text{var}(\hat{\gamma}_{RB}) = \text{var}(\hat{\gamma}_0) - E[\text{var}(\hat{\gamma}_0|d_r)]. \tag{10}$$

If $\hat{\text{var}}(\hat{\gamma}_0)$ is an estimator of $\text{var}(\hat{\gamma}_0)$ then an estimator of $\text{var}(\hat{\gamma}_{RB})$ is

$$\hat{\text{var}}(\hat{\gamma}_{RB}) = E[\hat{\text{var}}(\hat{\gamma}_0)|d_r] - \text{var}(\hat{\gamma}_0|d_r). \tag{11}$$

This estimator is the difference of the expectation of the estimated variance of the preliminary estimator over all consistent reorderings of the original data and the variance of the preliminary estimator over all consistent reorderings of the original data. Although these estimates are unbiased, they can result in negative estimates of the variance. For such a case a conservative approach is to set the estimate of $\text{var}(\hat{\gamma}_{RB})$ equal to $E[\hat{\text{var}}(\hat{\gamma}_0)|d_r]$.

## 4.3 Approximations through resampling

In the event there is a large number of sampling occasions and/or sample sizes, evaluating the Rao-Blackwell estimators may be computationally difficult as there will likely be a prohibitively large number of data reorderings to tabulate. An alternative approach is to base approximations for the Rao-Blackwell estimators on a Markov chain Monte Carlo (MCMC) resampling method. We have developed such a procedure based on the Metropolis algorithm. In the appendix we outline the resampling procedure and how to obtain approximations to

the Rao-Blackwell estimators with the procedure.

# 5 Population Quantity Estimators

As multiple samples are selected for the study, a mark-recapture model can be exploited to obtain estimates of the selection probabilities. Suitable plug-in estimators of population quantities can then be obtained based on these estimates. In this section we outline two plug-in estimators for the population total. We conclude with justifying a jackknife approach for estimating the variance of these estimators.

## 5.1 Population Total Estimators

The population total is defined to be $\tau = \sum_{i=1}^{N} y_i$ where $y_i$ is the response of interest of unit $i$. We consider two commonly used population total estimators, namely the Hansen-Hurwitz and Horvitz-Thompson estimators. For notational convenience we define $p^{(i)}$ to be the selection probability of unit $i$ (that is, $P_{ik} = p_j = p^{(i)}$).

### 5.1.1 Hansen-Hurwitz type estimator

Recall that $s_k$ represents the set of individuals selected for sample $k$, $k = 1, 2, ..., K$. A Hansen-Hurwitz (hereafter HH) type of plug-in estimator for the population total is

$$\hat{\tau}_{HH} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i \epsilon s_k} \frac{y_i}{\hat{p}^{(i)}} \tag{12}$$

where $\hat{p}^{(i)}$ is an estimate of $p^{(i)}$. It can be shown that the Rao-Blackwellized version of the HH-type estimator is

$$\hat{\tau}_{HH,RB} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i \epsilon s_k} \frac{y_i}{\hat{p}_{RB}^{(i)}} \tag{13}$$

where $\hat{p}_{RB}^{(i)}$ is the improved estimate of $\hat{p}^{(i)}$.

### 5.1.2 Horvitz-Thompson type estimator

Recall that $s = \cup_{k=1}^{K} s_k$ represents the set of individuals selected at least once over all sampling occasions. A Horvitz-Thompson (hereafter HT) type of plug-in estimator for the population total is

$$\hat{\tau}_{HT} = \sum_{i \epsilon s} \frac{y_i}{\hat{\pi}^{(i)}} \tag{14}$$

where $\pi^{(i)} = 1 - (1 - p^{(i)})^K$ is the inclusion probability of unit $i$ (that is, the probability that unit $i$ is selected at least once over all sampling occasions) and $\hat{\pi}^{(i)} = 1 - (1 - \hat{p}^{(i)})^K$ is an estimate of $\pi^{(i)}$. It can then be shown that the Rao-Blackwellized version of the HT-type estimator is

$$\hat{\tau}_{HT,RB} = \sum_{i \epsilon s} \frac{y_i}{\hat{\pi}_{RB}^{(i)}} \tag{15}$$

where $\hat{\pi}_{RB}^{(i)} = 1 - (1 - \hat{p}_{RB}^{(i)})^K$ is the improved estimate of $\hat{\pi}^{(i)}$.

We note here that in the classic survey sampling setting the sampling selection probabilities are known and hence the Horvitz-Thompson estimator is already a function of the minimal sufficient statistic, namely $(s, \underline{y}_s)$. In our study the selection probabilities are unknown and have estimators with preliminary and improved counterparts, hence the further improvement on the Horvitz-Thompson type estimator.

### 5.2 Variance estimators for population total estimators

Plug-in style variance estimators for the population totals are likely to be sensitive to the approximations of the inclusion probabilities. Hence, a well-known robust estimator for the variance of these estimators is suggested. We explore the use of the simple jackknife procedure suggested by Rust (1985). The method is outlined as follows.

Recall that $n = |s|$ is the number of individuals selected at least once for the study. Suppose $\hat{\tau}$ is an estimate of the population total based on all observations corresponding with $s$. Define $\hat{\tau}^{(i)}$ to be the estimate corresponding with the observations made on $s$ after removing unit $i \; \epsilon \; s$ and its corresponding observations from the study. The estimate of the variance of $\hat{\tau}$ is taken to be

$$\hat{\text{var}}(\hat{\tau}) = \frac{n-1}{n} \sum_{i \epsilon s} (\hat{\tau}^{(i)} - \hat{\tau})^2. \tag{16}$$

## 6 Empirical Study

There has been a rapidly growing interest in the use of mark-recapture methods for estimating the size of hidden populations; see Hook and Regal (1995) for a discussion on the subject. We therefore explore the new strategy via an empirical study of individuals at high-risk for HIV/AIDS in the Colorado Springs area (Darrow et al., 1999; Klovdahl et al., 1994; Rothenberg et al., 1995). The population data is based on Project 90, a prospective study funded by the Center for Disease Control and Prevention.[1] The population is summarized in Figure 1. Links between pairs of individuals indicate drug-using relationships and all links are reciprocated. The size of the population is 595 and the number of links in the population is 1458. As many empirical studies of injection drug-using populations can benefit from estimates of the rate of exchange of needles and other drug-using paraphernalia (Woodhouse

[1] for more details on Project 90, see https://opr.princeton.edu/archive/p90/

et al., 1994), the variable of interest in our study is the number of links in, also known as the density of, the population.

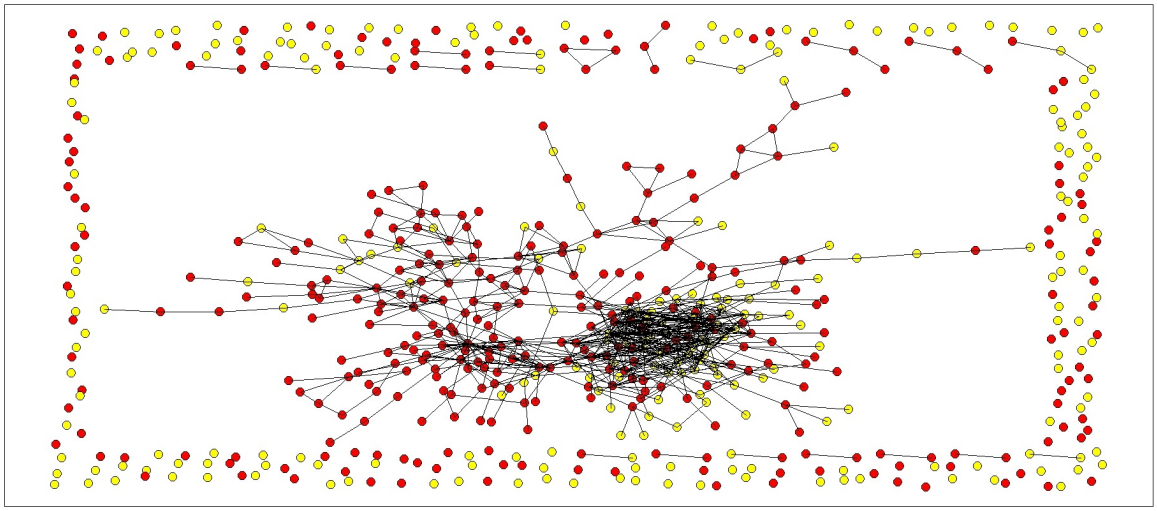

Figure 1: HIV/AIDS at-risk population (Klovdahl et al., 1994). Dark-coloured nodes indicate individuals who are injection drug-users and light-coloured ones indicate individuals who are non-injection drug-users. The size of the population is 595 and the number of links is 1458.

From a sampling perspective, there is a sense that the more an individual interacts with other individuals the more conspicuous they are. Therefore, in our study we base selection probabilities on the number of links emanating from individuals in that they are proportional to the square root of the number of these links (also known as the node-degree), after adding a constant. Figure 2 displays a histogram of the selection probabilities.

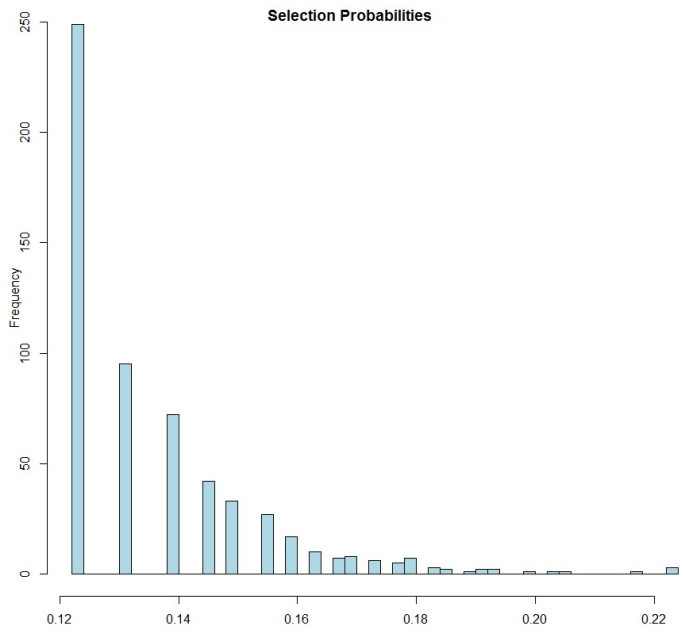


Figure 2: Histogram of selection probabilities for empirical study. The population is partitioned into twenty-three strata where members in the same stratum have equal node-degree.

## 6.1 Simulation Study Results

The simulation study is based on obtaining 2000 sets of three samples of individuals from the empirical population. Approximations to the improved versions of the preliminary estimators are based on the Markov chain resampling procedure outlined in the appendix, and 500 resamples were sufficient to accurately approximate the improved estimators as the acceptance rate is approximately 80%. The high rate of acceptance is due to the semi-uniform candidate distribution for selecting sample reorderings and the result that all consistent sample reorderings have equal probability of being selected in the empirical setting (see Expression (9)).

We base estimates of selection probabilities on the following mark-recapture estimators/models[2]:

[2]With respect to the LP estimator, the corresponding estimate of the selection probability is taken to be the ratio of the sum of the number of individuals selected for both of the first two sampling occasions and one to the sum of the number of individuals selected for the second sampling occasion and one. For all other estimators we make use of the "Rcapture" package (Baillargeon and Rivest, 2012) in the R program-

- LP: The Lincoln-Petersen estimator (Chapman, 1951), which is based on the $M_0$ assumption, corresponding with the first two samples,
- $M_0$ likelihood: the maximum likelihood estimator based on the $M_0$ assumption (Rivest and Baillargeon, 2007),
- Chao LB: Chao's $M_h$ lower bound estimator (Chao, 1987),
- Poisson2: the Poisson2 (using a Poisson model) estimator based on an $M_h$ assumption (Rivest and Baillargeon, 2007), and
- Darroch: Darroch's $M_h$ estimator (Darroch et al., 1993).

Table 3 gives the approximate expectation and variance of the population total estimators, as well as the ratio of the variances of the improved and preliminary estimators. With respect to the estimators based on the $M_0$ assumption, the $M_0$ maximum likelihood-based estimators are already a function of the corresponding minimal sufficient statistic $(n, C)$ for $(N, p)$ (Rivest and Baillargeon, 2007) and have naturally been Rao-Blackwellized. In contrast, the population total estimators corresponding with the Lincoln-Petersen estimator can be further improved. These estimators are also based on the $M_0$ assumption and it can be seen that, through Rao-Blackwellization, this is a competitive estimator. With respect to the estimators based on the $M_h$ assumption, the population total estimators corresponding with Chao's $M_h$ lower bound estimator found some improvement with the new strategy. In contrast, both of the other estimators have benefited immensely from Rao-Blackwellization.

ming language; the "closedp.0()" function provides the estimated average selection probability. We use the corresponding value as the estimated selection probabilities for all sampled individuals.

Table 3: Approximate expectation and variance of preliminary and improved population total estimators with ratio of improvement of Rao-Blackwellized estimators to their preliminary counterparts. Results are based on a three-sample study for the empirical population. The population total is 1458.

| Estimator | Expectation | Var., Preliminary | Var., Improved | Ratio |
|---|---|---|---|---|
| HH LP | 1419 | 138213 | 47439 | 0.343 |
| HT | 1426 | 109086 | 44445 | 0.407 |
| HH $M_0$ | 1447 | 50607 | 50607 | 1.000 |
| HT | 1448 | 47525 | 47525 | 1.000 |
| HH Chao LB | 1480 | 57386 | 53466 | 0.932 |
| HT | 1476 | 52797 | 50024 | 0.947 |
| HH Poisson2 | 1442 | 312015 | 59322 | 0.190 |
| HT | 1449 | 231624 | 52430 | 0.226 |
| HH Darroch | 1622 | 1167836 | 137211 | 0.118 |
| HT | 1603 | 864753 | 112833 | 0.130 |

Table 4 provides coverage rates and semi-length approximations based on nominal 95% confidence intervals and the Central Limit Theorem (CLT) corresponding with the population total estimators. For this simulation study it can be seen that the simple jackknife procedure has lent itself well for variance estimation. Also, the new strategy seems to give rise to improved estimators with corresponding confidence intervals that have higher coverage rates and an expected smaller width than their preliminary counterparts. Though a small number of negative estimates of the variance of the improved estimators are found when using the Poisson2 and Darroch models, it is found that the suggested conservative approach outlined in subsection 4.2 facilitated in obtaining confidence intervals with a reasonable range.

Table 4: Coverage rates (CR) and semi-lengths (SL) of confidence intervals based on CLT.

| Estimator | CR, preliminary | CR, improved | SL, preliminary | SL, improved |
|---|---|---|---|---|
| HH LP | 0.900 | 0.952 | 796 | 526 |
| HT | 0.908 | 0.949 | 712 | 501 |
| HH $M_0$ | 0.939 | 0.939 | 475 | 475 |
| HT | 0.850 | 0.850 | 341 | 341 |
| HH Chao LB | 0.953 | 0.955 | 505 | 494 |
| HT | 0.878 | 0.874 | 370 | 358 |
| HH Poisson2 | 0.882 | 0.962 | 1202 | 736 |
| HT | 0.860 | 0.930 | 996 | 592 |
| HH Darroch | 0.830 | 0.997 | 2446 | 1476 |
| HT | 0.812 | 0.994 | 2098 | 1270 |

We also explore the utility of the Rao-Blackwell strategy for obtaining improved versions of the population size estimators based on the mark-recapture models. Tables 5 and 6 summarize the performance of the estimators. Variance estimators are taken to be those accompanying the population size estimators obtained with the "Rcapture" package. In general, it can be seen that the results parallel those based on the population total estimators.

Table 5: Approximate expectation and variance of preliminary and improved population size estimators with ratio of improvement of Rao-Blackwellized estimators to their preliminary counterparts. Results are based on a three-sample study for the empirical population. The population size is 595.

| Estimator | Expectation | Var., Preliminary | Var., Improved | Ratio |
|---|---|---|---|---|
| LP | 593 | 20900 | 4967 | 0.238 |
| $M_0$ | 598 | 5374 | 5374 | 1.000 |
| Chao's LB | 611 | 6195 | 5756 | 0.929 |
| Poisson2 | 584 | 46836 | 6960 | 0.149 |
| Darroch | 648 | 166534 | 19200 | 0.115 |

Table 6: Coverage rates (CR) and semi-lengths (SL) of confidence intervals based on CLT.

| Estimator | CR, preliminary | CR, improved | SL, preliminary | SL, improved |
|---|---|---|---|---|
| LP | 0.893 | 0.920 | 273 | 143 |
| $M_0$ | 0.944 | 0.944 | 149 | 149 |
| Chao's LB | 0.963 | 0.961 | 162 | 156 |
| Poisson2 | 0.846 | 0.889 | 401 | 274 |
| Darroch | 0.826 | 0.996 | 899 | 533 |

# 7 Discussion

In this manuscript we have established a foundation for design-based inference of the fixed-population parameter vector when the population size is unknown and a Poisson sampling design with unknown selection probabilities is used for selection. We have also derived the complete and minimal sufficient statistic for such a case and have contrasted this statistic

with the usual minimal sufficient statistic in survey sampling. We have detailed the Rao-Blackwellized estimators corresponding with our approach to inference and hence estimators can now be functions of the minimal sufficient statistic. We have shown via an empirical study that one can expect a trade-off in relaxing the degree of the heterogeneity assumption with an increase in precision of estimators. In the appendix, we have also outlined a Markov chain resampling procedure that can be used to approximate the improved estimators when computation is not feasible.

In our study we relied on the jackknife approach to obtain estimators of the variance of the population total estimators. Future research on applying more sophisticated jackknife estimators when a small number of units in some strata are selected, like that presented in Kott (2001), would be worthwhile exploring. The authors are planning further work on this topic.

In the traditional survey sampling setting where the population size and selection probabilities are known, the minimal sufficient statistic is $d_R = ((s, \underline{y}_s))$ (Thompson and Seber, 1996). Recall in our study that we have shown that the minimal sufficient statistic is $d_R = ((s, \underline{y}_s), \underline{C})$ and therefore reflects on an attribute of the sampling design, namely unknown selection probabilities. Further investigation into how the minimal sufficient statistic changes over sampling designs, when the population size and selection probabilities are unknown, will make for interesting future work.

Expression (10) reveals that, in general, the greater the variability amongst the preliminary estimates corresponding with sample reorderings, the greater the expected improvement in the Rao-Blackwellized estimators. With the Poisson sampling design, the more a study is comprised of many samples and/or smaller sample sizes the more estimation will benefit from Rao-Blackwellization. Future work on how a similar Rao-Blackwellization strategy can contribute to analyses based on other commonly used sampling designs would be helpful.

Estimates of population quantities other than the population size will likely be required to

draw meaningful conclusions on populations like the hidden one we explored in the empirical simulation study. As there is a growing interest for studying such populations for which there is an absence of a full sampling frame, more research on how to obtain efficient estimates will be required. The methods used in this manuscript can possibly lend ideas for developing inference procedures that are based on the fixed-population model.

# References


Ahmad, M., Alalouf, S., and Chaubey, Y. P. (2000). Estimation of the population total when the population size is unknown. *Statistics & Probability Letters* **49,** 211–216.

Baillargeon, S. and Rivest, L.-P. (2012). *Rcapture: Loglinear Models for Capture-Recapture Experiments*. R package version 1.3-1.

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *The Indian Journal of Statistics, Series A* **31,** 441–454.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Wadsworth and Brooks/Cole, second edition.

Cassel, C.-M., Särndal, C.-E., and Wretman, J. H. (1977). *Foundations of inference in survey sampling*. Wiley New York.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43,** 783–791.

Chao, C.-T., Dryver, A. L., and Chiang, T.-C. (2011). Leveraging the rao-blackwell theorem to improve ratio estimators in adaptive cluster sampling. *Environmental and Ecological Statistics* **18,** 543–568.

Chapman, D. (1951). Some properties of the hypergeometric distribution with applications to zoological sample census. *University of California Publications in Statistics* **1,** 131–160.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88,** 1137–1148.

Darrow, W. W., Potterat, J. J., Rothenberg, R. B., Woodhouse, D. E., Muth, S. Q., and Klovdahl, A. S. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The colorado springs study. *Sociological Focus* **32,** 143–158.

Fattorini, L. (2006). Applying the horvitz-thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabiliites. *Biometrika* **93,** 269–278.

Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews* **17,** 243–264.

Kindahl, J. K. (1962). Estimation of means and totals from finite populations of unknown size. *Journal of the American Statistical Association* **57,** pp. 61–91.

Klovdahl, A., Potterat, J., Woodhouse, D., Muth, J., Muth, S., and Darrow, W. (1994). Social networks and infectious disease: The colorado springs study. *Social Science & Medicine* **38,** 79–88.

Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics* **4,** 521–526.

Rivest, L.-P. and Baillargeon, S. (2007). Applications and extensions of chao's moment estimator for the size of a closed population. *Biometrics* **63,** 999–1006.

Rothenberg, R. B., Potterat, J. J., Woodhouse, D. E., Darrow, W. W., Muth, S. Q., and Klovdahl, A. S. (1995). Choosing a centrality measure: Epidemiologic correlates in the colorado springs study of social networks. *Social Networks* **17,** 273–297. Social networks and infectious disease: HIV/AIDS.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics* **1,** 381–397.

Thompson, S. K. (2006). Adaptive web sampling. *Biometrics* **62,** 1224–1234.

Thompson, S. K. (2012). *Sampling.* Wiley Series in Probability and Statistics, New Jersey, third edition.

Thompson, S. K. and Seber, G. A. F. (1996). *Adaptive Sampling.* Wiley Series in Probability and Statistics, New York.

Woodhouse, D., Rothenberg, R., Potterat, J., Darrow, W., Muth, S., Klovdahl, A., Zimmerman, H., Rogers, H., Maldonado, T., and Muth, J. B. (1994). Mapping a social network of heterosexuals at high risk for hiv infection. *AIDS* **8,** 1331–1336.

# A Appendix

This appendix outlines the Markov chain resampling procedure with mathematical details that is used in the empirical study to obtain approximations to the Rao-Blackwellized estimators and corresponding estimates of the variance of these improved estimators.

## A.1 Candidate outcome distribution

We will first define an *outcome* as one that results in a specific sequence for which units are assigned to samples corresponding with a hypothetical sample reordering according to the steps outlined below. Suppose there are $G$ stratum. Repeat steps 1 and 2 once for each stratum $j = 1, 2, ..., G$.

Step 1: Suppose the number of unique units from stratum $j$ that are obtained for the final sample $s$ is equal to $n_j$. Assign each member to one of the samples completely at

random.

Step 2: For $k = 1, 2, ..., K$, let $l_{(k,j)}$ be the number of units from stratum $j$ that are in $s$ and that have not (possibly yet) been selected for sample $k$. Select a sample to receive an additional unit with probability proportional to $l_{(k,j)}$. Suppose the sample selected is $k^*$. Now, select a unit from stratum $j$ completely at random amongst those $l_{(k^*,j)}$ units not yet selected for sample $k^*$. Repeat this step a total of $C_j - n_j$ times, where $C_j$ is the total number of units originally obtained (counting repeat selections) from stratum $j$ over all sampling occasions.

**Claim**: With the aforementioned outcome selection procedure, all possible outcomes have equal probability of being selected.

**Proof**: Let $Q_j$ be the (uniform) probability of assigning individuals from stratum $j$ to the hypothetical samples as is done in step 1. If $C_j = n_j$, define $L_j = 1$. Otherwise, define $l_{(k,j,i)}$ to be the number of units from stratum $j$ in $s$ not selected for sample $k$ prior to part $i$ of step 2, for $i = 1, 2, ..., C_j - n_j$, and $k^*$ to be the sample selected at part $i$. Also define

$$L_j = \prod_{i=1}^{C_j - n_j} \left[ l_{(k^*,j,i)} / \sum_{k=1}^{K} l_{(k,j,i)} \times 1/l_{(k^*,j,i)} \right] = \prod_{i=1}^{C_j - n_j} \left[ 1 / \sum_{k=1}^{K} l_{(k,j,i)} \right]. \tag{17}$$

Then, the probability of selecting a specific outcome, $o^*$ say, under the resampling procedure is

$$P(o^*) = \prod_{j=1}^{G} \Big( Q_j L_j \Big). \tag{18}$$

Notice that this probability is uniform amongst all outcomes since $\sum_{k=1}^{K} l_{(k,j,i)}$ remains constant over all reorderings for each $i = 1, 2, ..., C_j - n_j$. This gives the claim. □

## A.2 Mapping of outcomes to sample reorderings and accept-reject ratio

Consider a sample reordering consistent with the minimal sufficient statistic. Let $f_{(j,i)}$ be the number of times unit $i$ in stratum $j$ is selected over all sampling occasions for this reordering. The total possible number of outcomes that correspond with this sample reordering is

$$\prod_{j=1}^{G}\Big\{\Big[\prod_{i\epsilon s\cap U_j}\binom{f_{(j,i)}}{1}\Big]\Big(\sum_{i\epsilon s\cap U_j}(f_{(j,i)}-1)\Big)!\Big\}, \tag{19}$$

where $U_j$ represents the units of stratum $j$. With the aforementioned outcome selection procedure, the candidate distribution is one that works through the outcome selection procedure so that the reordering corresponding with the sampled outcome is the candidate sample reordering. Recall that all sample reorderings consistent with the minimal sufficient statistic have the same probability of being selected in the empirical setting (see Expression (9)). Hence, the accept-reject aspect of the algorithm is based on the ratio of the number of outcomes that give rise to the most recently accepted sample reordering to that from the candidate sample reordering.

## A.3 Approximations to the Rao-Blackwellized estimates

With the aforementioned Markov chain resampling procedure, one can obtain an approximation to a Rao-Blackwellized estimator as follows. Suppose $\gamma$ is a population quantity or selection parameter to be estimated with a Markov chain of length $R$. The MCMC procedure commences with the original sample data so that it is in its stationary distribution and hence will remain in its stationary distribution at each iteration; let $\hat{\gamma}_0^{(0)}$ be the preliminary estimate of $\gamma$ obtained with the original data. Let $\hat{\gamma}_0^{(l)}$ be the preliminary estimate of $\gamma$ obtained with the most recently accepted sample reordering at iteration $l$. An enumerative

estimate of $\hat{\gamma}_{RB}$ is then

$$\tilde{\gamma}_{RB} = \frac{\sum_{l=0}^{R} \hat{\gamma}_0^{(l)}}{R+1}. \tag{20}$$

Similarly, let $\hat{\text{var}}(\hat{\gamma}_0^{(0)})$ be the estimate of the variance of $\hat{\gamma}_0$ obtained with the original data. Let $\hat{\text{var}}(\hat{\gamma}_0^{(l)})$ be the estimate of the variance of $\hat{\gamma}_0$ obtained with the most recently accepted sample reordering at iteration $l$. An enumerative estimate of $\hat{\text{var}}(\hat{\gamma}_{RB})$ is then

$$\begin{aligned} \tilde{\text{var}}(\hat{\gamma}_{RB}) &= \tilde{E}[\hat{\text{var}}(\hat{\gamma}_0) \mid d_r] - \tilde{\text{var}}(\hat{\gamma}_0 \mid d_r) \\ &= \frac{1}{R+1} \sum_{l=0}^{R} \hat{\text{var}}(\hat{\gamma}_0^{(l)}) - \frac{1}{R+1} \sum_{l=0}^{R} (\hat{\gamma}_0^{(l)} - \tilde{\gamma}_{RB})^2. \end{aligned} \tag{21}$$